System Dynamics Review
System Dynamics Review vol 28, No 3 (July-September 2012): 281–294
Published online 2 August 2012 in Wiley Online Library
(wileyonlinelibrary.com) **DOI**: 10.1002/sdr.1478

ARCHIVES

Partial-model testing as a validation tool for system dynamics (1983)

Jack B. Homer

System dynamicists have long emphasized the use of multiple data sources and multiple methods to estimate parameters and test models. Ideally, one should estimate parameters using data "below the level of aggregation of the model". For example, in a model of capital investment, one could estimate the length and order of the construction delay directly from data on the construction times for a large sample of relevant projects. Often, however, the needed data are not available. At the other end of the spectrum one can use "whole model estimation" in which the parameters are found by fitting the behavior of the full model to the available aggregate time series data. Whole model estimation, however, often suffers from identification problems. In this 1983 paper Jack Homer describes partial model testing, in which parameters are estimated within a subset of model structure rather than by calibration of the entire model. Homer illustrates with an example from his research on the adoption of new medical innovations, specifically, the cardiac pacemaker. He illustrates the partial model testing process with an important formulation representing how the clinical and research communities carry out and report follow-up data on the safety and efficacy of a medical innovation as it evolves.

The paper also illustrates the necessity of careful empirical work to collect new data. Debates over methods to estimate parameters are sterile without the data to implement them. In the context of medical innovation, it would have been plausible to assume that follow-up data evaluating pacemaker efficacy would grow smoothly over time as use expanded, though perhaps with a lag. Not content with such easy assumptions, however, Homer examined every article on pacing published in every issue of the relevant cardiology journals, from the creation of the pacemaker through the (then) present. The work, done years before the advent of the Internet and online databases, was painstaking and time consuming—the journals had to be searched and articles coded by hand. The payoff was a unique dataset documenting the actual dynamics of evaluative reporting. Rather than smooth growth, the data showed a pronounced oscillation in the publication of evaluative studies, even though other data, which Homer also assembled from original sources, showed smooth growth in adoption, clinical indications and pacemaker use. Homer's model generates the same oscillation endogenously, and the partial model tests provide robust estimates of the parameters governing the institutional processes (such as research and publication delays) and behavioral decision rules (such as the decision by researchers and clinicians to initiate new follow-up studies) that determine the existence, period, and amplitude of the cycle. The structure identified through this process was not only important to accurately model the evolution of the pacemaker, but has important policy implications still relevant today. All modelers should follow Homer's example and put in the hard work to generate, from primary sources, the data needed to estimate the important parameters and relationships in our models.

John Sterman

Homer J. 1983. Partial-model testing as a validation tool for system dynamics. In *Proceedings of the 1983 International System Dynamics Conference*. System Dynamics Society, Chestnut Hill, MA; 919–932.

Abstract

This paper discusses an approach to model refinement that involves testing the behavior of individual pieces of a model in response to empirical input data for comparison with empirical output data. Partial-model tests should be used for selecting formulations or estimating parameters only when appropriate case-specific or logical information is not available for this purpose. The smaller the model components used for partial-model testing, the more likely it is that the model will prove useful for anticipating events outside historical experience and the less likely it is that observed behavior will be incorrectly attributed to certain relationships or parameters.

Thus, from the standpoint of structural validity, partial-model testing is an improvement over whole-model testing for the purpose of structural adjustment. The paper presents a detailed example of partial-model testing in the context of a generic model of the evolving use of a new medical technology. Specifically, the technique is used for adjusting and validating a model subsystem that can explain why the reporting of clinical information on cardiac pacemakers has been marked by regular oscillations over time. Copyright © 2012 System Dynamics Society.

Syst. Dyn. Rev. **28**, 281–294 (2012)

Introduction

Validation of a system dynamics model can be viewed broadly as a process of demonstrating that both the structure and behavior of the model correspond to existing knowledge about the system under investigation (Forrester, 1961, Ch. 13; Forrester and Senge, 1980). Matching empirical reference mode data is frequently the focus of behavioral validation efforts, although the production of unexpected yet believable patterns of behavior can also enhance the perceived usefulness of a model (Forrester and Senge, 1980; Mass, 1991). Reproduction of empirical reference modes is most convincing when the model has been constructed and its parameters selected on the basis of detailed information concerning pieces of the system that are linked closely in time and space, rather than from correlations apparent in aggregate time-series data. Ideally, all elements of structure should be based on information at a level more detailed than that of the model itself (Graham, 1980). The more solid this foundation of data “below the level of aggregation of model structure” is, the more likely it is that the model will prove useful for anticipating events outside the historical experience and generating accurate insights for policymakers. Conversely, to the extent formulations and parameter values are picked with the sole aim of duplicating aggregate time-series data, the greater is the risk of producing a structure which breaks down outside of the historical context and leads to faulty conclusions.

Although one would like to formulate structures and estimate parameter values entirely on the basis of disaggregate case-specific data, this is often impossible in actual practice, where time and resource constraints limit the effort that can be devoted to background research. Models of "soft" systems, in which many of the quantities in question are not directly measurable, pose particularly severe difficulties as far as the collection of useful "micro" data is concerned. Of course, one can often use logic or knowledge gained from general experience to fill in the structural gaps left by empirical research; indeed, "educated guesses" are part and parcel of the model-building process. But even after case-specific research and modeler judgment are applied to construct a model, some small number of gaps may remain. These gaps may be of two types: (1) Alternative formulations for a given relationship may seem equally appealing, and (2) one may be unable to establish a reasonably narrow range of acceptable values for a given parameter. In such situations, one is thrown back upon the use of aggregate time-series data as a guide to structure. This use of time-series data means that a certain amount of curve-fitting will be done in order to select formulations or parameter values; the purity of a full *a priori* approach is thereby sacrificed. But this shortcoming need not deal the validation effort a fatal blow; it is possible to test and adjust a model without forcing or "fixing" the final result. Judicious use of the technique of partial-model testing is the key.

DOI: 10.1002/sdr

The technique: description, rationale, and strategy

A partial-model test involves simulating the behavior of a functional component of the model, which may be as small as a single equation, in response to empirical input data for comparison with empirical output data. The comparison with output data may be made by eye or with statistical methods, as is true of any other data-based testing scheme (Hamilton, 1980; Naill, 1973).[i] One begins with a guess as to which formulation or range of parameter values is likely to provide an acceptable fit to the historical data, where "goodness of fit" is defined relative to the model's purpose (Forrester, 1961). If the initial fit is not acceptable, the uncertain structure of interest is then adjusted until one finds a formulation or range of parameter values for which an acceptable fit is obtained. (If the component being tested contains *N* uncertain parameters, then one searches for a region in the relevant *N*-dimensional parameter space for which the fit is acceptable.) Note that formulations and parameter values will always be constrained to some extent by real-life considerations; all pieces of structure must make sense even if they are uncertain. For instance, logical considerations will often dictate that a parameter value not be negative or that it be less than some maximum value.

Partial-model testing should be viewed as preferable to whole-model testing for the purpose of selecting or estimating pieces of structure. This follows from the idea that one would like to make such choices in a way that is consistent with available information but does not inappropriately "fix" the final simulation results. Indeed, as the size of the structure and the number of uncertain parameters in any single test increase, the potential for structural misspecification or misattribution of behavior to particular parameters also increases. For example, if two uncertain time constants together determine the frequency of an observed oscillation, estimation of both parameters in a whole-model context could lead to offsetting estimation biases (Graham, 1980). Such misestimation could prove to be a critical flaw if the dominant real-life behavior has the potential of shifting from the original oscillation to some other mode in which one or both of the time constants still play a role. This problem could conceivably be avoided by estimating each uncertain parameter independently, in its own partial-model test. In general, the idea that uncertain formulations and parameter values should be established as independently as possible of the full model's structure suggests that the pieces of structure used in partial-model testing be as small as the available empirical data permit.[ii]

The component-by-component approach advocated here makes only partial use of all available data and structure in any given test. This means that a simulation of the whole model using parameter values established via partial-model testing will do no better and will generally do worse in terms of fit than a simulation in which a "full information" approach to parameter estimation is taken (Peterson, 1980). However, this is precisely what one would expect from a technique that seeks to give greater certainty to structure without "fixing" the final simulation results. The smaller the functional component being tested and the smaller the number of parameters adjusted to improve that component's response to input data, the closer one comes to the ideal of determining structure on the basis of focused considerations of closely-related phenomena rather than the full range of observed macro-behavior.

The fact that partial-model testing does not guarantee the desired behavior of the full model implies that some problems in formulation or parameter estimation may not be revealed until the full model is simulated. Indeed, in actual practice it seems that some

DOI: 10.1002/sdr

degree of adjustment and fine-tuning of the full model is inevitable and may even be the sign of a rich model—one in which the various components are coupled closely together. This suggests that partial-model testing might best be viewed as one step in an overall strategy of model development and testing which is intended to maximize both structural and behavioral validity. In the interest of structural validity, formulations and parameter values should be based on disaggregate case-specific data or other *a priori* considerations wherever possible, and partial-model testing used only as a secondary technique when the initial hypotheses fail to produce the desired result. Similarly, whole-model adjustments should be made only if partial-model tests have failed to reveal model inadequacies which become apparent when the full set of feedback loops is simulated intact. Not only does such a hierarchical approach to model-building maximize the likelihood that the model will be useful for examining events outside the historical context (since the structure is developed as independently as possible of the full range of observed behavior), but it may also be the best general approach to troubleshooting during model development. This is simply because it is easier to isolate the source of inappropriate behavior when it is seen in a small piece of structure than when it is seen in a larger and more complex structure. Thus, while "passing" a partial-model test does not guarantee the appropriateness of a given piece of structure, "failing" the test immediately identifies an area where improvement is needed and can prevent much of the head-scratching involved in troubleshooting a large, complex model.

Applying the technique: an example

Background

The remainder of this paper is devoted to presenting an example of partial-model testing in the context of a generic model of the evolving use of a new medical technology (Homer, 1983). This model was developed for the purpose of generating insights useful to government policymakers regarding the complex process by which new medical technologies are disseminated, improved, and controlled. The model focuses on the actions of physicians who recommend, use, and evaluate the technology, and manufacturers who are responsible for promoting and modifying the technology. The full model consists of approximately 150 equations, including over 30 levels and delays. It was developed over a three-year period and its structure based on an extensive review of the pertinent literature as well as in-depth case studies of two different medical technologies.[iii]

One of the cases studied was the implantable cardiac pacemaker, a sophisticated electronic device used for restoring or maintaining a normal heartbeat in certain patients who suffer from chronic heart rhythm problems (arrhythmias). Ever since the first implantation in 1960, the field of pacing has grown steadily, despite some early resistance by physicians to the unusual new device. As Figure 1 indicates, the number of patients with pacemakers ("recipients") in the U.S. increased from about two thousand in 1965 to over twenty thousand in 1970 to over 100,000 in 1975, and by 1980 stood at just over 350,000. This growth is only partially attributable to acceptance of the technique, which was essentially complete by 1971. Rather, the rapid growth of the 1970s is almost fully a reflection of expanding clinical indications for pacing, which came in response to several important technical developments that made the device both safer and more broadly applicable.

DOI: 10.1002/sdr

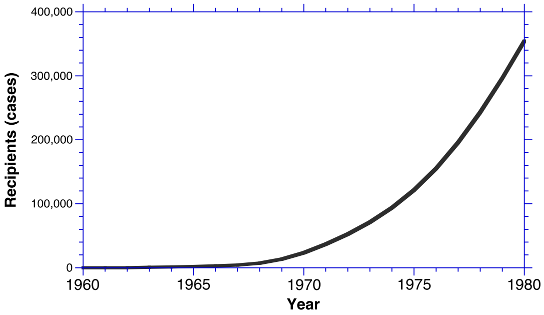

Fig 1. Patients with implanted pacemakers in the U.S. 1960–1980 (annual estimates). Source: Homer (1983, pp. 110–111)

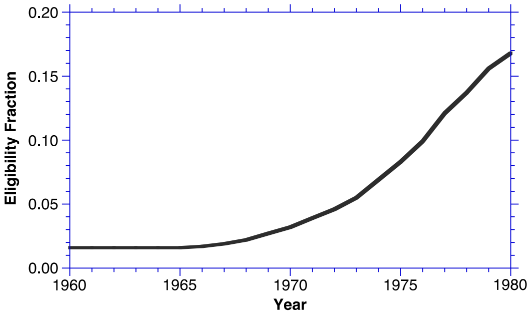

Fig 2. Fraction of arrhythmic patients deemed eligible for pacing (annual estimates). Source: Homer (1983, pp. 110–111)

Figure 2 shows how the fraction of heart rhythm patients considered eligible for pacing increased tenfold from less than two percent during the early 1960s to nearly seventeen percent by 1980. Originally, only patients with severe symptoms (dizziness, fainting, seizures) and at high risk of sudden death were considered appropriate candidates. But as physicians became bolder in their use of the device, they expanded the eligibility criteria to include certain less symptomatic (sometimes asymptomatic) patients as well as a whole class of patients with symptoms but with little risk of sudden death (Homer, 1983, Ch. 3).

Journal articles are a source of information that can affect the extent to which a new medical technology is used (Stross and Harlan, 1979; Manning and Denson, 1980; Banta *et al.*, 1981; Young, 1981). The articles with the greatest impact are those that appear in influential journals, describe clinical results in detail, and report on a large sample of cases. Several physicians, when interviewed, pointed to four journals which they consider to be most informative on the subject of pacing.[1] These journals were searched thoroughly for articles which appeared during the period 1960–1980 and which presented clinical follow-up data on the observed results of pacing for specified medical conditions. Several different kinds of information were gleaned from the fifty-six articles that satisfied these criteria, including the number of recipients reported per article. This information was used to construct Figure 3, which shows the total cases reported in the four journals on an annual basis (the "reporting rate"). Since larger sample sizes have greater impact, this time series may be considered to be an indication of how the influence of pacing articles changed over time.

The observed pattern of reporting is clearly oscillatory and appears to have a period of about five years, with an amplitude that is small at first but then becomes larger. This time series is striking, particularly because it is the only one of several pacing time series that were compiled that displays oscillations. Such oscillations are nowhere discussed in the literature. Indeed, very little has been written on the factors that underlie evaluation and publication decisions in general (relatedly: Levy and Sondik, 1978). The component of the model in which the reporting rate is generated therefore seems to be a prime candidate

[1] The four journals are: (1) *PACE* (*Pacing and Clinical Electrophysiology*), (2) *Circulation*, (3) *American Journal of Cardiology*, and (4) *American Heart Journal*. See Homer (1983, Appendix P3).

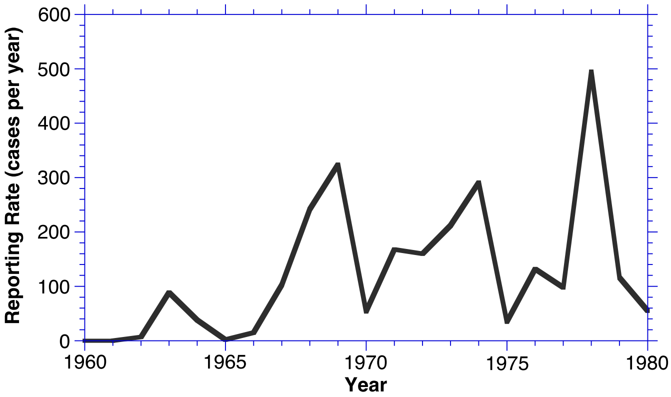

Fig 3. Paced patients reported in influential journals (annual). Source: Homer (1983, p. 135)

for partial-model testing: Can this component (whose structure is largely speculative) reproduce the observed oscillations, given inputs that are not themselves oscillatory?

Evaluation and reporting subsystem

Since little has been written about the processes of evaluation and reporting, the structure to be presented here is based primarily on inferences drawn from the case studies, particularly the pacemaker case study.[2] The goal was to make this structure as simple and transparent as possible by making only a small number of common-sense assumptions which seem generalizable beyond the specific context of the case studies. The resulting structure is shown in Figure 4. (An equation listing of the evaluation and reporting subsystem can be found in the Appendix.)

The Reporting Rate, as discussed previously, represents the publication in influential journals of clinical case information and is measured in cases published per year. Thus, Reports to Date represents the cumulative number of recipients whose procedures appear in the broadly recognized clinical literature on the technology. The Reporting Rate is simply a delayed version of the Evaluation Rate, which corresponds to the annual number of cases selected for analysis which do eventually make it into the journals.[3] The delay separating evaluation and reporting represents the total time required to write, submit, and publish an article on clinical outcomes. A comparison of submission and publication dates for pacemaker reports suggested that the Reporting Time generally exceeds one-half year and may even exceed two years, although one year is more typical. (This parameter can be estimated using partial-model tests, as discussed below.)

Physicians select patients for evaluation from the total pool of recipients, in response to a perceived need for more evaluative data on the technology. Both evaluators and journal editors will tend to turn their attentions to other topics as they become more certain of the technology's capabilities and limits. In other words, the fraction of recipients selected for evaluation (the Evaluation Fraction) will be responsive to the adequacy of existing

[2] The structure presented here is a bit simpler than that used in Homer (1983), which includes a "controversy" factor that was not of importance in the pacemaker case.

[3] Note the assumption that all evaluations are eventually published, which seems to ignore the process by which papers are screened for publication and many rejected. In fact, the screening process is represented in the model, but for the sake of simplicity has been placed at the stage of evaluation rather than reporting.

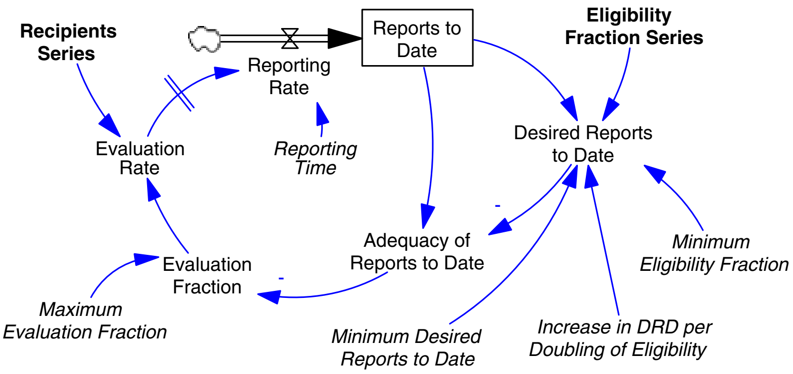
Recipients Series
Reporting Rate
Reports to Date
Eligibility Fraction Series
Evaluation Rate
Reporting Time
Desired Reports to Date
-
Adequacy of Reports to Date
Evaluation Fraction
-
Minimum Eligibility Fraction
Maximum Evaluation Fraction
Minimum Desired Reports to Date
Increase in DRD per Doubling of Eligibility

Fig 4. Evaluation and reporting subsystem. Source: Homer (1983, Ch. 5)

reports (Adequacy of Reports to Date); the fraction will be greatest (equal to the Maximum Evaluation Fraction) when no data are available and smallest when reports are seen as fully adequate. The relationship is represented analytically in the model as

$$\text{Evaluation Fraction} = (\text{Maximum Evaluation Fraction})(1 - \text{Adequacy of Reports to Date})^2$$

This nonlinear function of Adequacy of Reports to Date says that evaluators will be more sensitive to changes in the need for data when adequacy is low than when it is high. One can think of this curve as reflecting the relative number of physicians who choose to investigate the technology in response to a given level of need: When the adequacy of reports is high, only a small number of researchers with special commitment will continue to evaluate the technology, while the bulk of physicians will only enter the fray if uncertainty is high and the journals are especially anxious to print new data on the technology. In the pacemaker case, a national survey showed that fully a third of implanting physicians have published articles on pacing (Parsonnet and Manhardt, 1977), although probably most of them no more than once or twice. On the basis of this information, the Maximum Evaluation Fraction was set at 0.3 (30% per year), which says that virtually all of these physicians would respond in the event that no data on pacing existed.

The Adequacy of Reports to Date is the ratio of existing Reports to Date to Desired Reports to Date. The desired level will be no smaller than the current stock of information, but may be expected to increase as the Eligibility Fraction increases. One may look at this from a statistical perspective: As eligibility expands, the population being assessed grows proportionally, so that a larger sample is needed to achieve a given level of confidence in the results. More concretely, eligibility criteria widen through the inclusion of new subsets of patients, and since outcomes may be highly dependent on the particular subset being considered (Weinstein and Stason, 1977; National Academy of Sciences, 1978; Fineberg and Hiatt, 1979; Banta *et al.*, 1981), such an expansion in scope generally implies a need for new data. The pacemaker case provides a good example of this phenomenon: In the early 1960s, articles focused on results for patients who had severe symptoms and were at high risk of sudden death; in the late 1960s, the focus switched to a newly-included subset of patients with less risk of sudden death; in the early 1970s, articles concentrated on patients with "sick sinus syndrome", whose risk of sudden death is negligible; and in

DOI: 10.1002/sdr

the late 1970s, the focus of articles shifted to the benefits of the "prophylactic" pacing of patients who have few or no symptoms associated with their arrhythmias but are at some risk of sudden death (such as certain heart attack victims).[4] Indeed, one can largely associate each fluctuation in Figure 3 with a period of renewed clinical research activity brought on by the application of pacing to a new subset (or subsets) of arrhythmic patients.

The following nonlinear relationship was selected to describe the connection between the Eligibility Fraction and the Desired Reports to Date[5]:

$$\begin{aligned}\text{Desired Reports to Date} = \text{MAX}\big(&\text{Reports to Date}, \text{MNDRD} + \text{IDRDEL}\\ &*\log_2(\text{Eligibility Fraction}/\text{Minimum Eligibility Fraction})\big)\end{aligned}$$

where MNDRD is Minimum Desired Reports to Date and IDRDEL is Increase in Desired Reports to Date per Doubling of Eligibility.

This formulation was suggested both by theoretical considerations (sampling theory suggests the required sample size should rise as the square root of the population under consideration) and by a comparison of the historical values of Reports to Date and the Eligibility Fraction in the pacemaker case. This comparison is shown in Figure 5. Each point on the graph corresponds to one year between 1960 and 1980.[6] If one assumes that the adequacy of reports was for the most part high during this period—that is, that the response to needed data has always been relatively rapid and unhindered (this description does seem to fit the case of pacing well)—then reports to date can be used as a rough proxy for desired reports to date, and the observed relationship can be used to estimate the two parameters MNDRD and IDRDEL. The logarithmic curve superimposed on the empirical data in Figure 5 shows that a good fit is obtained when MNDRD = 100 cases and IDRDEL = 700 cases.

Since aggregate time-series data were used to derive these two estimates, what has just been described is actually an example of partial-model parameter estimation (even if the piece of structure is only a single equation). Thus, some partial-model testing of a small piece of the evaluation and reporting subsystem was found to be necessary before the whole subsystem could be tested. In general, the strategy of testing components as small as the data permit can lead to such a sequence of partial-model tests, in which tests of one piece of the model are best performed after another uncertain piece has been tested and adjusted. In the present instance, the evaluation and reporting loop can now be tested without further adjustment of the parameters MNDRD and IDRDEL.

Simulation testing

In this section, the use of partial-model testing to estimate the order and length of the delay between evaluation and reporting will be demonstrated.[7] The time series on Recipients

[4] See Appendix P3 of Homer (1983) on the pacemaker literature.

[5] Desired Reports to Date is never less than Reports to Date and has an absolute minimum value of MNDRD, which corresponds to a situation in which Eligibility Fraction equals Minimum Eligibility Fraction. In the pacemaker case, the Eligibility Fraction was equal to its minimum value of .016 during the period 1960–1965; physicians were extremely cautious in their selection of patients to receive the device during these early years.

[6] The eligibility fraction series was plotted in Figure 2. The reports to date series was calculated by accumulating the reporting rate series, which was plotted in Figure 3.

[7] Hamilton (1980) discusses various single-equation estimation techniques for this problem when data on both the direct input to the delay (Evaluation Rate) and the output to the delay (Reporting Rate) are available. But since data on the evaluation rate were not available, a larger piece of structure—in fact, the whole evaluation and reporting subsystem—was required for estimation purposes.

DOI: 10.1002/sdr

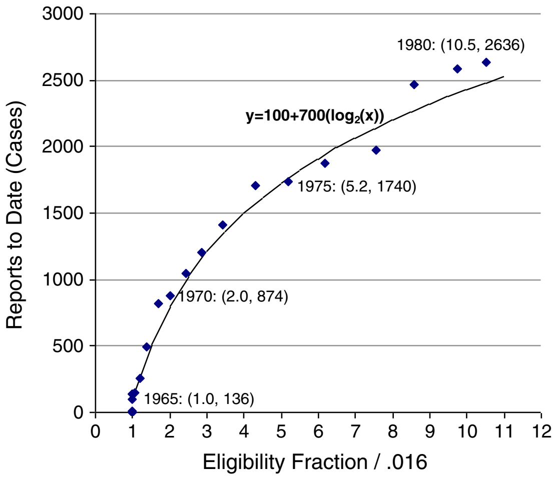

Fig 5. Eligibility Fraction versus Reports to Date

and Eligibility Fraction, shown in Figures 1 and 2, respectively, were used as inputs in this testing, and the simulated Reporting Rate compared with the historical data shown in Figure 3. In the full medical technology model, both recipients and the eligibility fraction are generated endogenously, and the reporting rate can feed back, through various channels, to affect these variables. Because of this feedback, a successful partial-model test does not guarantee equal success in the context of the full model. But as discussed previously, this is precisely the sacrifice made in exchange for the greater structural validity gained when structures are estimated using the smallest possible number of assumptions and aggregate time series in any given test.

In the interest of simplicity of structure, a first-order delay was initially used for testing purposes. The reporting time was adjusted between one-half and two years (the range suggested by prior information) until a simulated reporting rate with period and amplitude characteristics most like those of the historical time series was found. Figure 6 shows the best result, obtained when the reporting time was set to 1.25 years. The simulated fluctuations are

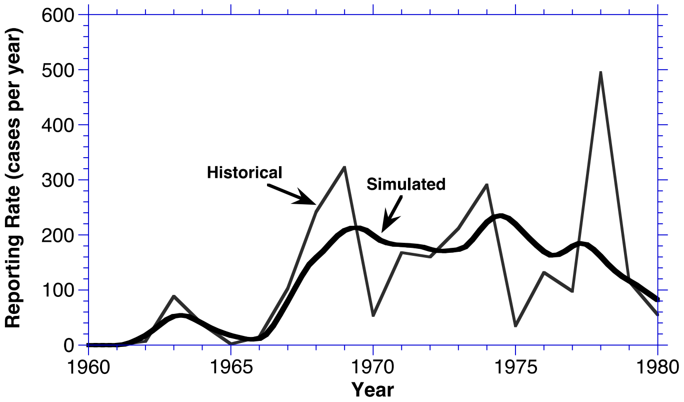

Fig 6. Partial-model test, first-order delay ($T = 1.25$)

neither as strong nor as regular as those seen in the empirical data. An astute system dynamicist might guess that the first-order delay does not introduce enough of a phase lag into the subsystem's negative loop to bring out this loop's oscillatory potential to a sufficient degree. Increasing the order of the delay therefore makes sense as the next step in partial-model testing. But in the present instance, one need not argue for a higher-order delay, say, third order, on these behavioral grounds alone. The fact that the reporting delay actually represents the multi-stage process of writing, submitting, and publishing an article means that a structural defense for the implied increase in model complexity exists as well. Indeed, rarely will a physician be in the position of evaluating patients one day and seeing his or her results published the next, as would happen frequently if the process were really a first order one.

The move to a third-order delay does, in fact, improve the subsystem's response considerably. Figure 7 shows the best result, again obtained when the reporting time was set to 1.25 years. Not only are the frequency and growing amplitude of the simulated oscillation true to the original, but the timing matches as well. Surely this test serves to strengthen one's confidence in the evaluation and reporting subsystem.

At this juncture, it is appropriate to explain the observed oscillations as they are produced in the model. Fundamental is the fact that the reporting rate can grow to exceed that rate required to just satisfy the need created by expanding eligibility. The source of this information glut is the delay between evaluation and reporting, which enables a backlog of not-yet-published evaluations to accumulate. If the publication process were fully centralized, this would not cause a problem, because the backlog would be taken into account. But since there are many evaluators and generally several journals to which they can send evaluations, several reports on the same subject may appear simultaneously even if only one of them is necessary to satisfy the current need for information.

Given this background on how over-reporting can occur, the following story can be told for the pacemaker case: As the eligibility fraction increased roughly exponentially over time, the demand for reports also increased, roughly linearly as the logarithmic function would suggest. In response to the increasing demand for evaluative data, reports were submitted and published at a rate that turned out to somewhat exceed current needs and which led to periods during which the cumulative stock of reports appeared adequate. But the eligibility fraction continued to climb, so the complacency would eventually wear off and a whole new round of evaluation and reporting would begin. As the number of recipients who could be evaluated grew, the response to a given need for data increased, which explains the growing amplitude of oscillation.

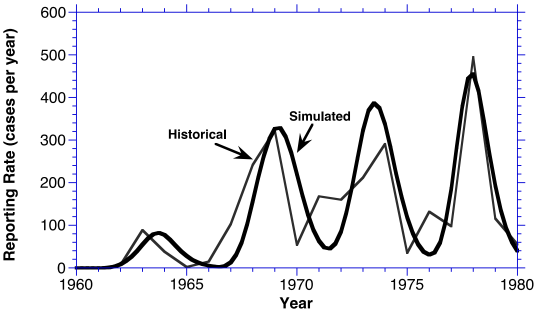

Fig 7. Partial-model test, third-order delay ($T = 1.25$)

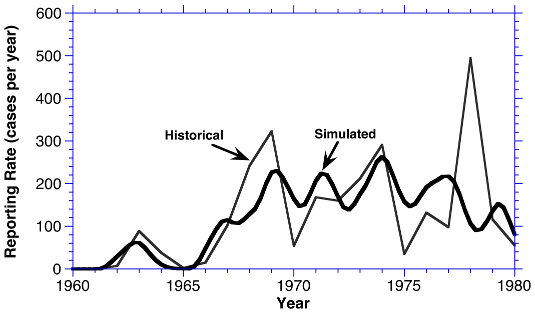

Fig 8. Partial-model test, third-order delay, shorter delay time ($T = 0.5$)

DOI: 10.1002/sdr

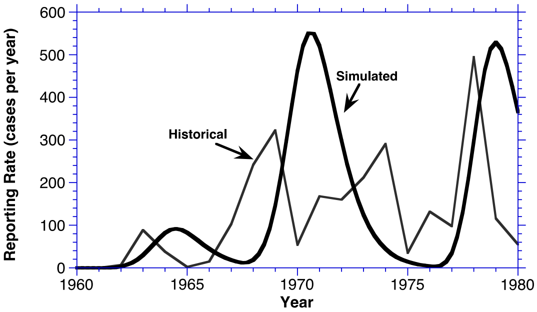

Fig 9. Partial-model test, third-order delay, longer delay time ($T = 2.0$)

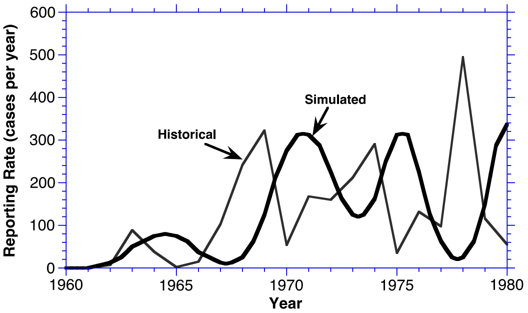

Fig 10. Whole-model simulation of the reporting rate. Source: Homer (1983, p. 431)

It is useful to examine how sensitive the simulated oscillation is to the delay time. Figure 8 shows the result of reducing the delay time to one-half year, while Figure 9 shows the result of increasing the delay time to two years. Clearly, as the delay time increases, both the period and amplitude of oscillation also increase. The longer the reporting delay is, the greater the potential for excess evaluation and reporting activity, so the greater the amount of overshoot and the longer it takes for the resulting complacency to be worn away by expanding eligibility. From these and other partial-model tests, it appears that the period of oscillation is roughly proportional to (about four times) the reporting delay time, a high degree of sensitivity by any measure.

Having settled upon formulations and parameter values that perform well in the partial-model context, it is desirable to examine the simulated reporting rate in the context of the full model. Figure 10 presents a comparison of whole-model simulation results with the empirical data. Even with all of the model's feedback loops intact—specifically, with recipients and eligibility fraction being generated endogenously—the amplitude and period of oscillation are still reproduced faithfully. However, this simulation does show that the timing of oscillation is affected; the simulated reporting rate lags the historical rate by about a year throughout the twenty-year period. Apparently, the slight differences between the simulated and historical values of recipients and eligibility (and these differences are quite small) are enough so that the oscillations are triggered at slightly different times in the whole-model test than in the partial-model test.

This raises an interesting point. Since it has been shown that the evaluation and reporting subsystem *can* produce the correct timing of oscillations under the controlled conditions of a partial-model test, one is led to the conclusion not that something is wrong with the model, but rather that, in real life, the precise timing of the oscillations is sensitive to random disturbances that can affect the evaluation and reporting process. Thus, although the period and amplitude of oscillation appear to be robust features of the system's behavior, timing seems to be a more unpredictable matter. It would be difficult to make this point convincingly without performing a partial-model test that reproduces the historical timing of oscillation.

Conclusion

This paper has outlined and demonstrated an approach to pinning down uncertain formulations and parameter values. In actual modeling practice, one frequently does not

DOI: 10.1002/sdr

have access to as much disaggregate case-specific or logical information as would be necessary to specify a complex model in its entirety. In such circumstances, some compromising of the ideal *a priori* approach to model construction is necessary, leading to a certain amount of curve-fitting. This paper has argued that there are better and worse ways of going about this, given the desire to avoid "fixing" the final result. The key is doing partial-model testing instead of whole-model testing whenever possible. More generally, the suggested strategy is to set formulations and parameter values on the basis of tests of model components that are as small as the available time series data permit. This approach not only minimizes the likelihood of incorrectly attributing behavior to "innocent" (or only marginally involved) relationships or parameters, but can also lend greater efficiency to the process by which a model is improved.

Endnotes

i. Sterman (1984) later described the value of Theil inequality statistics for evaluating the historical fit of system dynamics models, and Oliva (1995) later developed a Vensim^{TM} module for calculating these summary statistics along with simulation output. These statistics are of particular value when it is not readily apparent to the eye whether the model is producing outputs with means or variances different from those in the historical data.

ii. This theme was further developed in later years by Oliva (2003), who, in a discussion of automated calibration techniques, sums it up as follows: "Working with small calibration problems reduces the risk of the structure being forced into fitting the data, increases the efficiency of the estimation (estimators with smaller variances), and concentrates the differences between observed and simulated behavior in the piece of structure responsible for that behavior."

iii. A later article (Homer, 1987) gives an overview of the two case studies, model structure and base run results, compared to historical data.

References

Banta HD, Behney CJ, Willems JS. 1981. *Toward Rational Technology in Medicine*. Springer: New York.

Fineberg HV, Hiatt RH. 1979. Evaluation of medical practices: the case for technology assessment. *New England Journal of Medicine* **301**: 1086–1091.

Forrester JW. 1961. *Industrial Dynamics*. MIT Press: Cambridge, MA. Reprinted by Pegasus Communications: Waltham, MA.

Forrester JW, Senge PM. 1980. Tests for building confidence in system dynamics model. In *System Dynamics, TIMS Studies in the Management Science*, Vol. **14**. North-Holland: New York; 209–228.

Graham AL. 1980. Parameter formulation and estimation in system dynamics models. In *Elements of the System Dynamics Method*, Randers J (ed.). MIT Press: Cambridge, MA. Reprinted by Pegasus Communications: Waltham, MA.

Hamilton MS. 1980. Estimating lengths and orders of delays in system dynamics models. In *Elements of the System Dynamics Method*, Randers J (ed.). MIT Press: Cambridge, MA. Reprinted by Pegasus Communications : Waltham, MA.

Homer JB. 1983. A dynamic model for analyzing the emergence of new medical technologies. PhD dissertation, MIT, Cambridge, MA.

DOI: 10.1002/sdr

Homer JB. 1987. A diffusion model with application to evolving medical technologies. *Technological Forecasting and Social Change* **31**: 197–218.

Levy RI, Sondik EJ. 1978. Decision-making in planning large comparative studies. *Annals of the NY Academy of Sciences* **304**: 441–457.

Manning PR, Denson TA. 1980. How internists learned about cimetidine. *Annals of Internal Medicine* **92**: 690–692.

Mass NJ. 1991 (written 1981). Diagnosing surprise model behavior: a tool for evolving behavioral and policy insights. *System Dynamics Review* **7**(1): 68–86.

Naill RF. 1973. The discovery life cycle of a finite resource: a case study of U.S. natural gas. In *Toward Global Equilibrium*, Meadows DL, Meadows DH (eds). Wright-Allen Press: Cambridge, MA. Reprinted by Pegasus Communications: Waltham, MA.

National Academy of Sciences. 1978. The evaluation of equipment-embodied technology. In *A Study of the Diffusion of Equipment-Embodied Technology*. Government Printing Office: Washington, DC.

Oliva R. 1995. A Vensim module to calculate summary statistics for historical fit. D-4584, System Dynamics Group, MIT, Cambridge, MA.

Oliva R. 2003. Model calibration as a testing strategy for system dynamics models. *European Journal of Operational Research* **151**: 552–568.

Parsonnet V, Manhardt M. 1977. Permanent pacing of the heart: 1952 to 1976. *The American Journal of Cardiology* **39**: 250–256.

Peterson DW. 1980. Statistical tools for system dynamics. In *Elements of the System Dynamics Method*, Randers J (ed.). MIT Press: Cambridge, MA. Reprinted by Pegasus Communications: Waltham, MA.

Sterman JD. 1984. Appropriate summary statistics for evaluating the historical fit of system dynamics models. *Dynamica* **10**(Winter): 51–66.

Stross JK, Harlan WR. 1979. The dissemination of new medical information. *Journal of the American Medical Association* **241**: 2622–2624.

Weinstein MC, Stason WB. 1977. Foundations of cost-effectiveness analysis for health and medical practices. *The New England Journal of Medicine* **296**: 716–721.

Young DA. 1981. Communications linking clinical research and clinical practice. In *Biomedical Innovation*, Roberts EB, *et al*. (eds.). MIT Press: Cambridge, MA.

Appendix: evaluation and reporting subsystem equations used for partial-model testing

Active equations

Adequacy of Reports to Date = Reports to Date / Desired Reports to Date

Desired Reports to Date = MAX (Reports to Date, Minimum Desired Reports to Date + Increase in DRD per Doubling of Eligibility * 1.443 * LN(Eligibility Fraction / Minimum Eligibility Fraction))

<note: $\log_2(e) = 1.443$>

Evaluation Fraction = Maximum Evaluation Fraction * $(1 - \text{Adequacy of Reports to Date})^2$

Evaluation Rate = Recipients * Evaluation Fraction

Increase in DRD per Doubling of Eligibility = 700 cases

Maximum Evaluation Fraction = 0.3 per year

Minimum Desired Reports to Date = 100 cases

Minimum Eligibility Fraction = 0.016

Reporting Rate = DELAY3 (Evaluation Rate, Reporting Time) * (1 – Switch for first order delay) + DELAY1 (Evaluation Rate, Reporting Time) * Switch for first order delay

Copyright © 2012 System Dynamics Society

DOI: 10.1002/sdr

Reporting Time = 1.25 years
Reports to Date = INTEG(Reporting Rate, 0)
Switch for first order delay = 0

Historical data series

Eligibility Fraction = Eligibility Fraction Series(Time)
Eligibility Fraction Series ([(1960,0)–(1980,0.2)], (1960,0.016), (1961,0.016), (1962,0.016), (1963,0.016), (1964,0.016), (1965,0.016), (1966,0.017), (1967,0.019), (1968,0.022), (1969,0.027), (1970,0.032), (1971,0.039), (1972,0.046), (1973,0.055), (1974,0.069), (1975,0.083), (1976,0.099), (1977,0.121), (1978,0.137), (1979,0.156), (1980,0.168))
Historical Reporting Rate = Historical Reporting Rate Series (Time)
Historical Reporting Rate Series ([(1960,0)–(1980,600)], (1960,0), (1961,0), (1962,7), (1963,89), (1964,38), (1965,2), (1966,15), (1967,103), (1968,242), (1969,324), (1970,54), (1971,168), (1972,160), (1973,212), (1974,291), (1975,35), (1976,132), (1977,98), (1978,495), (1979,116), (1980,55))
Recipients = Recipients thousands Series (Time) * 1000
Recipients thousands Series ([(1960,0)–(1980,400)], (1960,0), (1961,0), (1962,0.2), (1963,0.7), (1964,1.3), (1965,2.1), (1966,3), (1967,4.3), (1968,7.2), (1969,13.5), (1970,23.6), (1971,36.8), (1972,52.7), (1973,71.2), (1974,93.8), (1975,121.6), (1976,155.2), (1977,195.8), (1978,243.2), (1979,296.3), (1980,354.6))

Simulation control parameters

FINAL TIME = 1980
INITIAL TIME = 1960
SAVEPER = TIME STEP
TIME STEP = 0.03125